PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(10): 891~899
www.pibb.ac.cn

高通量 RNA 甲基化测序数据处理与分析研究进展 *

刘 恋 [1)] 张绍武 [1)**] 孟 佳 [2)] 陈润生 [1, 3)]

([1)] 西北工业大学自动化学院，信息融合技术教育部重点实验室，西安 710072；
[2)] 西交利物浦大学生物科学系，吴江太湖新城研究院，苏州 215123；[3)] 中国科学院生物物理研究所，北京 100101)

摘要 随着高通量测序技术快速发展，MeRIP-seq (methylated RNA immunoprecipitation sequencing) 测序技术开启了 RNA 表观遗传学研究新局面，能够在全基因组范围内描述 RNA 甲基化．从 MeRIP-seq 高通量数据中挖掘 RNA 甲基化模式，有助于揭示 mRNA 甲基化在调控基因表达、剪切等方面所发挥的潜在功能，有效指导癌症的干预治疗．本文从 MeRIP-seq 测序原理出发，较全面地综述 MeRIP-seq 数据处理和分析方法研究现状，并对其所面临的计算问题进行讨论和展望．

关键词 MeRIP-seq 测序，数据处理与分析，RNA 甲基化，表观遗传
学科分类号 Q5，Q6，Q7 **DOI**: 10.16476/j.pibb.2015.0078

表观遗传学，包括组蛋白共价修饰(covalent histone modification)、DNA 甲基化修饰(DNA methylation)、RNA 甲基化修饰(RNA methylation)、基因组印记(genomic imprinting)、基因沉默(gene silencing)、RNA 编辑(RNA editing)及非编码 RNA (noncoding RNA)等，是指在核苷酸序列不发生改变的情况下，生物表型或基因表达发生了稳定的可遗传变化[1]．RNA 甲基化作为表观遗传学研究的重要内容之一，是指发生在 RNA 分子上不同位置的甲基化修饰现象，6-甲基腺嘌呤(N^6-methyladenosine，m^6A)和 5-甲基胞嘧啶(C^5-methylcytidine，m^5C)是真核生物中最常见的两种 RNA 转录后修饰．RNA 甲基化在调控基因表达、剪接、RNA 编辑、RNA 稳定性、控制 mRNA 寿命和降解等方面可能扮演重要角色．相对于 DNA 甲基化，RNA 甲基化更加复杂、种类繁多、普遍存在于各种高级生物中[2-4]．由于缺乏有效检测手段，相关研究多局限于非编码 tRNA 和 rRNA，或小部分编码转录片段[1]，且多数 RNA 甲基化功能未知．

随着高通量测序技术的发展[5]及一些 RNA 甲基化功能的发现[6-11]，人们开始关注 RNA 甲基化研究．尤其 MeRIP-seq (methylated RNA immunoprecipitation sequencing)高通量测序技术的出现，能够高效精确检测全转录组不同的 RNA 甲基化，奠定了 RNA 甲基化研究基础．如何有效处理和分析 MeRIP-seq 技术生成的海量数据，是成功发现 RNA 甲基化机理及功能的关键．

本文较全面介绍 MeRIP-seq 测序原理、数据处理及分析基本流程、关键方法、现有算法软件，重点讨论 MeRIP-seq 数据处理和分析过程中所面临的挑战．

1 MeRIP-seq 技术测序原理

MeRIP-seq 技术将甲基化 DNA 免疫共沉淀(methylated DNA immunoprecipitation，MeDIP)技术[12]、RNA 结合蛋白免疫共沉淀(RNA immunoprecipitation，RIP)技术和 RNA 测序(RNA sequencing，RNA-seq)技术[13]组合起来，高精度地检测全基因组(或全转录组)范围内的 RNA 甲基

* 国家自然科学基金资助项目(91430111, 61473232, 61401370, 61170134).
** 通讯联系人.
Tel: 029-88431308, E-mail: zhangsw@nwpu.edu.cn
收稿日期：2015-03-23，接受日期：2015-07-01

化．MeRIP-seq 技术采用免疫共沉淀方法，即甲基化 RNA 特异性抗体与被随机打断的 RNA 片段进行孵育，抓取有甲基化修饰的片段进行测序；同时需要平行测序一个对照(control)样本，对照样本用于消除抓取带有甲基化片段过程中的背景．然后将免疫共沉淀(IP)样本和对照样本中的序列片段对比(或定位)到参考基因组 / 转录组上，检测 RNA 甲基化位点．对照样本测量对应 RNA 的表达量，本质上是 RNA-seq 数据．图 1 为 MeRIP-seq 技术检测 m^{6}A RNA 甲基化过程示意图．

MeDIP-seq 和 ChIP-seq 测序技术均是将免疫共沉淀与测序相结合．MeRIP-seq 技术主要应用于 RNA 甲基化研究，而 ChIP-seq、MeDIP-seq 主要应用于 DNA 甲基化研究．MeRIP-seq 技术要求必须有对照样本，而 MeDIP-seq 和 ChIP-seq 技术对于对照样本没有要求．表 1 为 MeRIP-seq、MeDIP-seq 和 ChIP-seq 三种测序技术对比．

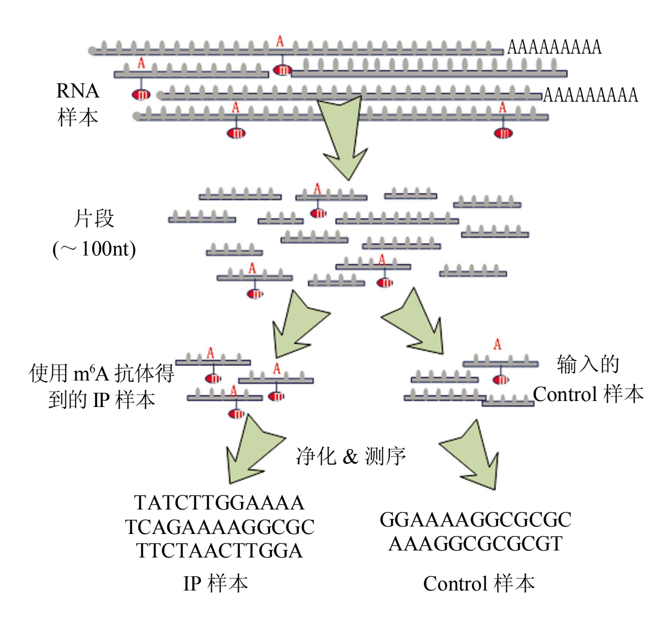


Fig. 1 The work flow of detecting m^{6}A RNA methylation using MeRIP-seq technology

图 1 MeRIP-seq 技术检测 m^{6}A RNA 甲基化过程

Table 1 Comparison of MeRIP-seq, MeDIP-seq, ChIP-seq sequencing technologies

表 1 MeRIP-seq、MeDIP-seq、ChIP-seq 三种测序技术对比

	ChIP-seq	MeDIP-seq	MeRIP-seq
研究对象	化学修饰	化学修饰	化学修饰
分子	DNA	DNA	RNA
比对器	非拼接	非拼接	拼接
特征	蛋白质绑定位点或峰	CpG 岛	甲基化位点或峰
量化	相对量(与绝对量线性相关)	相对量(与绝对量线性相关)	相对量(与绝对量不相关)
差异分析	仅需要免疫沉淀样本	仅需要免疫沉淀样本	需要免疫沉淀样本和对照样本
模体	双链	双链	链特异性
处理流程	读段定位 → 标签平移序列延伸 → Peak 富集分析 → 显著性分析	一般不做峰检测	在外显子上进行读段定位 → 每个滑窗中对 IP 和 control 样本的 reads 数泊松建模 → 比较 2 个泊松分布的均值，确定 Peaks → Peaks 的显著性分析
代表软件	MACS[14]，CisGenome[15]	Batman[16]，MeQA[17]	exomePeak[18]，MeRIP-PF[19]

2 MeRIP-seq 测序文库制备和测序平台数据输出

本小节将针对 Illumina/Solexa 测序平台，介绍 MeRIP-seq 测序文库制备及测序平台数据输出．

2.1 MeRIP-seq 测序文库制备

MeRIP-seq 测序文库制备过程如下：首先从样本细胞组织中分离出 RNA，考虑到总 RNA 中含有大量的 rRNA 序列，因此需要结合不同的方法去除其中的 rRNA．对于真核生物而言，常采用 Poly(T)寡核苷酸提取出带 Poly(A)的 RNA 去除 rRNA；而对不含 Poly(A)尾的转录本序列以及存在部分降解的总 RNA 样本而言，需要试剂盒去除 rRNA，从而得到除 rRNA 外的全部 RNA，然后将提取出的 RNA 随机打断．MeRIP-seq 技术对带有甲基化修饰的片段(IP 样本)进行测序时，需要平行对一个对

照样本(Control 样本)进行测序，其 IP 样本和 Control 样本的片段选择方法主要有以下 2 种：a．将打断的 RNA 片段分成两份，一份直接用于制备 Control 样本的 cDNA 文库，另一份采用抗 m^6A 抗体与被打断的 RNA 进行孵育，抓取带有 m^6A 修饰的片段，用于制备 IP 样本的 cDNA 文库．由于测序得到的结果不以所有 RNA 片段为背景，称这样得到的 IP 样本和 Control 样本是非成对的(unpair)，在进行数据处理时需先对 Control 样本进行处理．b．取两份相同的 RNA 进行打断，其中一份所有的 RNA 片段都进行测序，作为 Control 样本，另一份采用抗 m^6A 抗体抓取带有 m^6A 修饰的片段进行测序作为 IP 样本．由于测序得到的 IP 样本背景为当前测序得到的 Control 样本，称这样得到的 IP 样本和 Control 样本是成对的(pair)，可直接用于数据处理.

获取测序片段后(包括 IP 样本测序片段和 Control 样本测序片段)，用随机引物和反转录酶从 RNA 片段合成双链 cDNA．然后，对合成的 cDNA 进行末端修复并在 3′端加“A”，使用特定测序接头(adapter)连接 cDNA 片段两端，从而得到用于测序的 cDNA．通常情况下，为了得到更高的测序效率，一般采用电泳切胶法获取一定长度的 cDNA，再对其进行 PCR 扩增，得到所需的 cDNA 文库[20].

2.2 测序平台数据输出

将制备好的测序文库放入测序平台的各通道(lane)，通过桥式扩增，形成数以亿计的簇，开始测序．测序时，将 4 种聚合酶加入到单分子阵列中，每个被加入荧光标记的核苷释放出相对应的荧光．测序仪通过捕获荧光标记核苷酸所释放的荧光信号，利用计算机软件确定测得的碱基及顺序，根据测序顺序连成读段(read/fragment)，输出以 FASTQ 格式记录读段序列及测序质量分数.

在 FASTQ 文件中，每 4 行为一个读段，其中第 1 行以“@”开头，后面是 reads 的 ID 以及其他信息，第 2 行为测序得到的 read 的碱基序列，第 3 行以“+”开头，跟随着该 read 的名称(一般与 @ 后面的内容相同)，但有时可以省略，而“+”一定不能省，第 4 行代表 reads 的质量[21].

3 MeRIP-seq 测序数据处理

MeRIP-seq 技术主要用于 mRNA 甲基化检测，其测序数据处理主要包括读段定位、峰检测(peak calling)、差异甲基化检测及剪接异构体层次的相关处理．图 2 为 MeRIP-seq 测序数据处理流程.

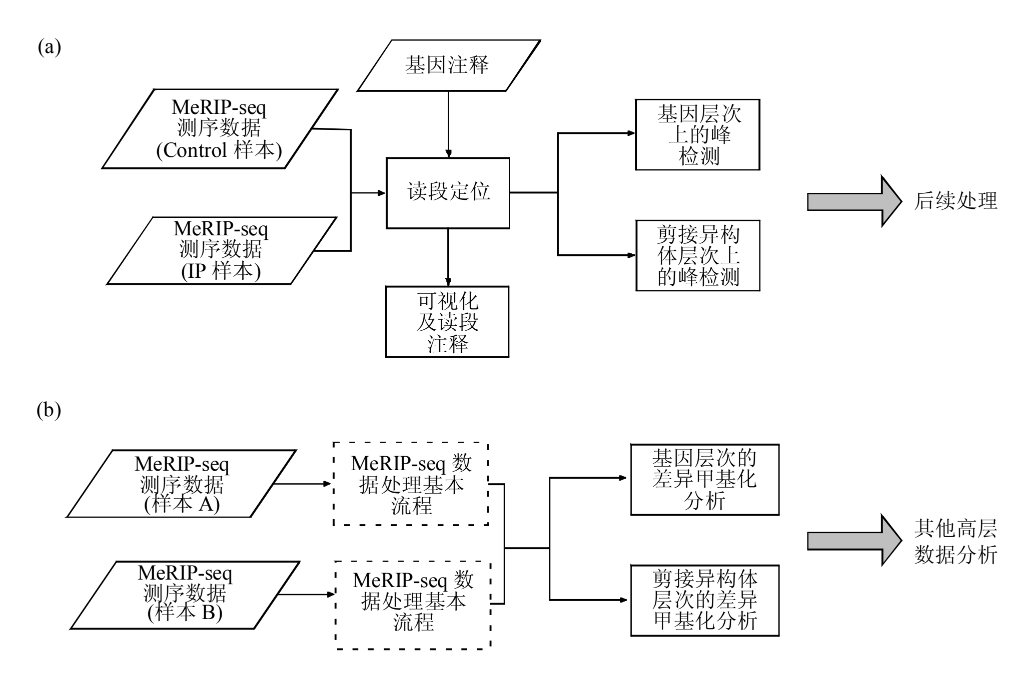


Fig. 2 The process of treating the MeRIP-seq data

图 2 MeRIP-seq 测序数据处理流程

(a) 单样本 MeRIP-seq 测序数据处理. (b) 双样本 MeRIP-seq 数据比较分析流程.

3.1　读段定位

获得 Control 及 IP 两样本测序数据后，首先对读段数据进行预处理(如将测序质量较差的读段过滤)，然后将两个样本的所有读段序列映射(mapping)定位到参考基因组上，这是后续数据处理和分析的基础．目前，RNA 数据的读段定位算法主要采用以下三种技术[22]：空位种子索引(spaced-seed indexing)、Burrows-Wheeler 转换 (Burrows-Wheeler transform，BWT)、Smith-Waterman 动态规划[23]．空位种子索引算法基本原理：将读段切成片段，形成种子片段，从中选取一部分作为种子建立索引，然后利用查找、延伸等方法来定位读段．其代表软件包括 MAQ[24]、ZOOM[25]、RMAP[26]．BWT 算法基本原理：通过 B-W 转换对参考基因组进行一次有规律的重新排序并建立索引，然后利用查找和回溯定位等方法进行读段定位．在查找过程中，可以利用碱基替代来实现允许的错配．其代表软件包括 Bowtie[27]、BWA[28]、SOAP2[29]．Smith-Waterman 动态规划算法基本原理：利用初始条件和迭代关系计算两条序列所有可能的比对分值，对相同位点加分，不同位点减分．采用空隙惩罚机制处理片段中存在的间隙，并将结果存放于一个矩阵中，利用动态规划方法回溯寻找最优比对结果．其代表软件有 BFAST[30]、SHRiMR[31]．

MeRIP-seq 测序数据实际上是一种 RNA 读段数据，读段定位时需要进行拼接定位，且读段定位中会面临跨越两个外显子结合区域的定位问题．为解决此问题，人们采用以下三种方法进行 RNA 读段定位：a．基于已知剪接点的比对定位．该方法在已知基因注释信息基础上实现，剪接点在已知接合区域数据库中可检测到．此类方法不能确定新的剪接点．代表性软件工具包括 SpliceSeq[32]、SAMMate[33]．b．从头拼接比对定位．此方法不需已知的注释信息，且允许新剪接点的检测．代表性软件工具包括 MapSplice[34]、SpliceMap[35]．c．使用注释信息进行从头拼接的比对定位．代表性软件工具包括 TopHat[36]、STAR[37]．TopHat 软件首先采用 Bowtie 比对非拼接的读段，然后采用 Maq 组装已比对的读段形成序列的岛；在岛屿序列中，TopHat 根据之前未映射的读段、可能的标准供体以及接受位点来确定剪接点．

读段定位后，通常采用 SAM[38]或 BAM 文件存储．BAM 格式是对 SAM 文件的压缩，可以将 SAM 格式压缩到接近原来的 20%．SAMTools[38]、BEDTools[39]、IGV[40]为 SAM 和 BAM 文件常用处理软件．

3.2　峰检测算法

IP 样本中甲基化位点抓取的读段较多，将其映射到参考基因组上，会在甲基化位点附近形成一个读段富集区(enrichment region)或者一个“峰(peak)，因而甲基化富集点检测算法称之为峰检测(peak calling)算法[41]．峰检测过程中，经常遇到两种比较特别的读段：一种为同一个读段可映射到基因组的多个位置上，称之为“多映射读段(multimapping reads)”；另一种为一些完全相同的读段，称之为“复制读段(duplicated reads)”，该类读段可能是由 PCR 扩增引起的．对于“多映射读段”，常采用下面 2 种方法处理：a．在不同位置根据周围区域情况按比例分配；b．完全删除这种读段，这是最简单并且最有效的方法[42]．对于“复制读段”，采用 SAMTools[38]处理．

MeRIP-PF[19]和 exomePeak[18]是目前检测 MeRIP-seq 数据读段富集区的两个主要工具．MeRIP-PF 首先将 IP 样本数据及对照样本数据映射到参考基因组上，并把参考基因组分割成 25 bp 的固定窗口，通过比对该窗口上 IP 样本和 Control 样本的读段(read)数目，确定 m^6A 甲基化区．但该 MeRIP-PF[19]以固定窗口分割参考基因组，对于跨窗口的“峰”及跨外显子的“峰”不能有效地处理，假阳性较高．exomePeak[18]采用 Przyborowski[43]和 Wilenski[43-44]方法比较两个泊松分布的均值(或 *C*-test)，对特定基因外显子集合进行峰检测，可检测跨越外显子连接区域的峰，该方法可有效解决转录丰度问题．由于基因剪接异构体的多样性，exomePeak 算法没有考虑转录复杂性，所涉及的诸如平移(shifting)、延伸(extension)、平滑(smoothing)、检测等计算操作相对直接简单．尽管 exomePeak 算法目前存在这些不足，但该算法仍可以较好地检测 RNA 甲基化位点，并对其进行注释．

3.3　差异甲基化检测

基于 MeRIP-seq 数据进行差异甲基化检测，有助于确定 2 种实验 / 显性条件(如正常和癌症)下的 mRNA 表观遗传调控差异．ChIP-Seq 数据与 MeRIP-seq 数据的差异甲基化检测有其本质区别．在 ChIP-Seq 数据中，由于 DNA 总数在两种情况下(加刺激、未加刺激)是相同的，那么修饰 DNA 分子的百分比与其数量保持相同的变化趋势，因此无论使用相对量(百分比)还是绝对量，其差异是一致

的．但在 MeRIP-seq 数据中，由于 mRNA 差异表达影响，MeRIP-seq 数据的背景(如 mRNA 转录丰度)差异较大．有可能同时出现“过甲基化(hypermethylation)”和“甲基化 RNA 总量下降”情况，如图 3 所示．在 DNA 甲基化中，未加刺激的情况下，3 个 DNA 分子中有 2 个被修饰，而加刺激情况下，3 个 DNA 分子中只有 1 个被修饰．在修饰的 DNA 分子质量下降的同时，其百分比也是下降的．但在 RNA 甲基化中，未加刺激的情况下，4 个 RNA 分子中有 2 个被甲基化，而在加刺激情况下，仅有 1 个 RNA 分子，且被修饰．即相对于未加刺激情况下的 RNA 甲基化，加刺激情况下的 RNA 甲基化数量虽然减少了，但其 RNA 甲基化百分比却增加．图 3 表明：由于 DNA 总量保持不变，甲基化 DNA 总量和其在总 DNA 中的相对量保持相同的变化趋势；由于 RNA 总量可能变化，甲基化 RNA 总量和相对量的变化可能完全不同．另外，图 3 中所示带有甲基化的 RNA 在加刺激中的总量虽然下降，可是其相对量却上升，表明了一种过甲基化现象，同时 RNA 表达量下调了．

exomePeak[18]工具包含差异甲基化区域检测功能，其检测原理基于超几何测试计算两种情况下的峰值富集显著性差异，且与一般情况下的 ChIP-Seq 和 RNA-seq 计算的绝对峰值差异不同．

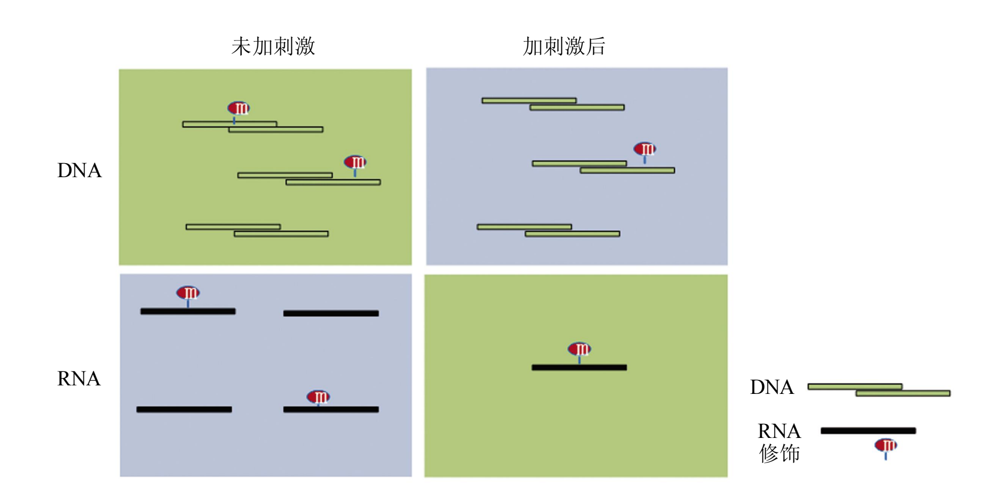


Fig. 3　The difference of DNA and RNA differential methylation

图 3　DNA 与 RNA 差异甲基化区别

4　MeRIP-seq 数据处理面临的生物信息学挑战

MeRIP-seq 技术为 RNA 表观遗传学开启了新的研究领域，但数据分析及处理方法的发展滞后于实验技术的进步，现有 DNA 甲基化数据分析及处理方法不能直接用来分析 RNA 甲基化数据，急需在以下几方面发展有效的计算方法，分析 MeRIP-seq 高通量 RNA 甲基化数据．

4.1　甲基化位点预测

与 ChIP-Seq 数据类似，基于 MeRIP-seq 数据的 RNA 甲基化位点预测需要消除背景读段分布噪声，如 GC 含量、映射能力、抗体非特异性结合、局部拷贝数变异等因素引起的实验误差和测序误差．ChIP-Seq 数据的背景偏差相对较小，其转录因子或 DNA 甲基位点预测不需要对照样本，仅通过估计邻居基因组区域的背景就可实现 DNA 甲基化位点预测[45-46]．与此相反，由于 mRNA 片段转录丰度变化较大及其在 3′和 5′端的衰减，MeRIP-seq 数据的背景读段分布变化非常大，必须通过对照样本测量背景转录丰度．因此 MeRIP-seq 数据甲基化位点预测需要检测相对于对照样本转录丰度的 IP 样本“富集峰(peak enrichment)”．因而，mRNA 甲基化位点检测与常用的 DNA 甲基化位点检测有本质上区别．

另外，当 RNA 甲基化位点处于外显子连接区

附近时，“峰”将跨越外显子连接区．因此 RNA 甲基化位点预测算法需要确定跨越 2 个或多个外显子的“峰”，否则，当采用现有诸如用于 ChIP-Seq 数据的 MACS[14]峰检测算法时，会错误检测出多个孤立“峰”．

虽然 exomePeak[18]能够实现跨越外显子连接的 RNA 甲基化位点检测，但 exomePeak 并没有完全解决上述 MeRIP-seq 所存问题．由于 exomePeak 通过泊松模型计算读段数目，没有考虑生物学差异，会遗漏过离散的读段．因此，需要发展新的 RNA 甲基化位点检测算法，以更加准确地进行 RNA 甲基化位点检测．

4.2　基因剪接异构体层次上 mRNA 甲基化预测

众所周知，高等真核生物中，通过可变性剪接相同的基因会被转录成不同的异构体(isoform)[47]，在基因剪接异构体上也会发生 RNA 甲基化．当 IP 样本中的一个峰处于异构体共享外显子时，进行 RNA 甲基化位点检测前，需对峰读段进行去卷积运算，确定每个异构体的相对贡献．另外，还需要确定对照样本中异构体表达的数量及它们的相应丰度．总之，如何应用 RNA-seq 对照数据预测不同异构体甲基化位点，是 MeRIP-seq 数据分析中一个迫切需要解决的挑战性问题．

4.3　基因及其剪接异构体层次上的 mRNA 差异甲基化预测

不同实验条件下，ChIP-Seq 数据的背景(基因组 DNA)通常非常相似，而由于 mRNA 的差异表达，MeRIP-seq 数据的背景(mRNA 转录丰度)差异较大．因而，现有适合于 ChIP-Seq 数据的差异分析算法[48]，不能直接用来比较两个 IP 样本中的读段数．需要研究包括相应 RNA-seq 对照样本的新计算框架，比较富集峰的相对数量．另外，针对某一转录异构体，需要研究有效的算法检测其差异甲基化．

4.4　基于分子网络的 RNA 甲基化功能注释

RNA 甲基化可通过调控基因表达而实施重要生物学功能，但 RNA 甲基化如何调控基因、究竟有哪些生物学功能，目前缺乏深入研究．我们可通过整合其他组学数据、构建与 RNA 甲基化相关的分子网络，采用相关的分子动态信息及网络分析方法[49-53]，研究 RNA 甲基化的基因调控机制及其所发挥的生物学功能．但如何与其他组学数据整合、如何构建 RNA 甲基化分子网络及如何挖掘分析也是目前急需解决的挑战性问题．

除上述迫切需要解决的问题之外，RNA-seq 分析中诸如多种读段、转录水平上的测序变化及比对偏差等因素，对于 MeRIP-seq 甲基化峰的检测同样重要，而 ChIP-Seq 数据分析方法中不需要考虑这些因素．另外，发展一些有效的方法将 RNA 甲基化数据与其他组学数据进行整合，深入研究 RNA 甲基化机理及其生物学功能也是生物信息学今后的一个重要研究方向．因而迫切需要发展新的针对 MeRIP-seq 数据的分析方法和计算工具解决上述问题，促进表观转录组学这一新兴领域的快速发展．

5　总结与展望

RNA 甲基化在调控基因表达、剪接、RNA 编辑、RNA 稳定性、控制 mRNA 的寿命和降解等方面可能扮演重要角色，其甲基化机理、位点预测和差异表达研究，有助于进一步揭示细胞发育、疾病等生物学现象，帮助药物研发者设计出能够调节基因表达、杀死或控制疾病细胞的小分子．本文从 MeRIP-seq 高通量测序技术出发，首先介绍此技术测序原理，在技术特征和数据处理流程方面与 MeDIP-seq、ChIP-seq 2 种高通量测序技术进行了对比，然后对 MeRIP-seq 高通量测序数据的读段定位、峰检测、差异甲基化检测及剪接异构体等相关处理方法进行归纳总结，最后，对 RNA 甲基化位点检测、剪接异构体层次上的甲基化位点检测、RNA 差异甲基化分析及基于分子网络的 RNA 甲基化功能注释所面临的生物信息学挑战问题进行了展望．希望本文能够对正在或即将采用 MeRIP-seq 实验进行科学研究的学者和 MeRIP-seq 高通量数据处理研究者提供参考．

参　考　文　献

[1] Fu Y, He C. Nucleic acid modifications with epigenetic significance. Current Opinion in Chemical Biology, 2012, **16**(5): 516-524

[2] Desrosiers R, Friderici K, Rottman F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. Proc Natl Acad Sci USA, 1974, **71**(10): 3971-3975

[3] Harris R A, Wang T, Coarfa C, *et al*. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature Biotechnology, 2010, **28**(10): 1097-1105

[4] Dubin D T, Taylor R H. The methylation state of poly A-containing-messenger RNA from cultured hamster cells. Nucleic Acids Research, 1975, **2**(10): 1653-1668

[5] Meyer K D, Saletore Y, Zumbo P, *et al*. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop

codons. Cell, 2012, **149**(7): 1635-1646

[6] Dominissini D, Moshitch-Moshkovitz S, Salmon-Divon M, *et al*. Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. Nature Protocols, 2013, **8**(1): 176-189

[7] Meyer K D, Jaffrey S R. The dynamic epitranscriptome: N6-methyladenosine and gene expression control. Nature Reviews Molecular Cell Biology, 2014, **15**(5): 313-326

[8] Liu J, Yue Y, Han D, *et al*. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nature Chemical Biology, 2013, **10**(2): 93-95

[9] Ping X L, Sun B F, Wang L, *et al*. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell Research, 2014, **24**: 177-189

[10] Jia G, Fu Y, Zhao X, *et al*. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nature Chemical Biology, 2011, **7**(12): 885-887

[11] 宋述慧, 李语丽, 于 军. RNA 中 6- 甲基腺嘌呤的研究进展. 遗传, 2013, **35**(12): 1340-1351
Song S H, Li Y L, Yu J. Hereditas, 2013, **35**(12): 1340-1351

[12] Weber M, Davies J J, Wittig D, *et al*. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature Genetics, 2005, **37**(8): 853-862

[13] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2009, **10**(1): 57-63

[14] Feng J, Liu T, Qin B, *et al*. Identifying ChIP-seq enrichment using MACS. Nature Protocols, 2012, **7**(9): 1728-1740

[15] Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotechnology, 2008, **26**(11): 1293-1300

[16] Down T A, Rakyan V K, Turner D J, *et al*. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature Biotechnology, 2008, **26**(7): 779-785

[17] Huang J, Renault V, Sengenes J, *et al*. MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. Bioinformatics, 2012, **28**(4): 587-588

[18] Meng J, Cui X, Rao M K, *et al*. Exome-based analysis for RNA epigenome sequencing data. Bioinformatics, 2013, **29**(12): 1565-1567

[19] Li Y, Song S, Li C, *et al*. MeRIP-PF: An easy-to-use pipeline for high-resolution peak-finding in MeRIP-Seq data. Genomics, Proteomics & Bioinformatics, 2013, **11**(1): 72-75

[20] 孙 磊, 张 林, 刘 辉. 基于 RNA-Seq 的长非编码 RNA 预测. 生物化学与生物物理进展, 2012, **39**(12): 1156-1166
Sun L, Zhang L, LIU H. Prog Biochem Biophys, 2012, **39**(12): 1156-1166

[21] Cock P J, Fields C J, Goto N, *et al*. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 2010, **38**(6): 1767-1771

[22] 王 曦, 汪小我, 王立坤, 等. 新一代高通量 RNA 测序数据的处理与分析. 生物化学与生物物理进展, 2010, **37**(8): 834-846
Wang X, Wang X W, Wang L K, *et al*. Prog Biochem Biophys, 2010, **37**(8): 834-846

[23] 杨 烨, 刘 娟. 第二代测序序列比对方法综述. 武汉大学学报: 理学版, 2012, **58**(5): 463-470
Yang Y, Liu J. J Wuhan Univ (Nat.Sci.Ed.), 2012, **58**(5): 463-470

[24] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 2008, **18**(11): 1851-1858

[25] Lin H, Zhang Z, Zhang M Q, *et al*. ZOOM! Zillions of oligos mapped. Bioinformatics, 2008, **24**(21): 2431-2437

[26] Smith A D, Xuan Z, Zhang M Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics, 2008, **9**(1): 128-136

[27] Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, **10**(3): R25.1-R25.10

[28] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 2009, **25**(14): 1754-1760

[29] Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 2009, **25**(15): 1966-1967

[30] Homer N, Merriman B, Nelson S F. BFAST: an alignment tool for large scale genome resequencing. PloS One, 2009, **4**(11): e7767

[31] Rumble S M, Lacroute P, Dalca A V, *et al*. SHRiMP: accurate mapping of short color-space reads. PLoS Computational Biology, 2009, **5**(5): e1000386

[32] Ryan M C, Cleland J, Kim R, *et al*. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. Bioinformatics, 2012, **28**(18): 2385-2387

[33] Xu G, Deng N, Zhao Z, *et al*. SAMMate: a GUI tool for processing short read alignments in SAM/BAM format. Source Code for Biology and Medicine, 2011, **6**(1): 2-13

[34] Wang K, Singh D, Zeng Z, *et al*. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research, 2010, **38**(18): e178-e178

[35] Au K F, Jiang H, Lin L, *et al*. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Research, 2010, **38**(14): 4570-4578

[36] Kim D, Pertea G, Trapnell C, *et al*. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 2013, **14**(4): R36

[37] Dobin A, Davis C A, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, **29**(1): 15-21

[38] Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, **25**(16): 2078-2079

[39] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, **26**(6): 841-842

[40] Robinson J T, Thorvaldsdóttir H, Winckler W, *et al*. Integrative genomics viewer. Nature Biotechnology, 2011, **29**(1): 24-26

[41] Valouev A, Johnson D S, Sundquist A, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods, 2008, **5**(9): 829-834

[42] Meng J, Cui X, Liu H, *et al*. Unveiling the dynamics in RNA epigenetic regulations//BIBM. 2013 IEEE International Conference on Bioinformatics and Biomedicine. Shanghai: BIBM, 2013: 139-144

[43] Krishnamoorthy K, Thomson J. A more powerful test for comparing two Poisson means. Journal of Statistical Planning and Inference, 2004, **119**(1): 23-35

[44] Przyborowski J, Wilenski H. Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder. Biometrika, 1940, **31**(3-4): 313-323

[45] Zhang Y, Liu T, Meyer C A, *et al*. Model-based analysis of ChIP-Seq (MACS). Genome Biol, 2008, **9**(9): R137

[46] Kharchenko P V, Tolstorukov M Y, Park P J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology, 2008, **26**(12): 1351-1359

[47] Pan Q, Shai O, Lee L J, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 2008, **40**(12): 1413-1415

[48] Xu H, Wei C L, Lin F, *et al*. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics, 2008, **24**(20): 2344-2349

[49] Zhang X, Zhao J, Hao J K, *et al*. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Research, 2014, **43**(5): e31

[50] Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. Journal of Theoretical Biology, 2014, **362**: 35-43

[51] Chen L, Liu R, Liu Z P, *et al*. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Scientific Reports, 2012, **2**: 342

[52] Liu R, Wang X, Aihara K, *et al*. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Medicinal Research Reviews, 2014, **34**(3): 455-478

[53] Wang J, Huang Q, Liu Z P, *et al*. NOA: a novel network ontology analysis method. Nucleic Acids Research, 2011, **39**(13): e87

Recent Progress and Challenges in High Throughput RNA Methylation Sequencing Data Analysis*

LIU Lian[1)], ZHANG Shao-Wu[1)**], MENG Jia[2)], CHEN Run-Sheng[1,3)]

([1)] *Key Laboratory of Information Fusion of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an* 710072, China;
[2)] *Department of Biological Sciences, XJTLU-WTNC Research Institute, Xi'an Jiaotong-Liverpool University, Suzhou* 215123, China;
[3)] *Institute of Biophysics, Chinese Academy of Sciences, Beijing* 100101, China)

Abstract With the rapid development of high-throughput sequencing technologies, the emerging of methylated RNA immunoprecipitation sequencing (MeRIP-seq) technology makes it possible to detect RNA epigenetic modifications in a large scale, which allows transcriptome-wide profiling of RNA methylation. Mining the patterns of global mRNA methylation from these MeRIP-seq data can help reveal the potential functional roles of these mRNA methylations in regulating gene expression, splicing, RNA editing and RNA stability, effectively guiding the therapeutic intervention of cancer. Here, the principle of MeRIP-seq sequencing was first introduced. Then, the recent progress of the processing and analysis of MeRIP-seq data were comprehensively discussed. In the end, the computational problems and challenges faced in the process of MeRIP-seq data processing were also summarized.

Key words MeRIP-seq sequencing, data processing and analysis, RNA methylation, epigenetics
DOI: 10.16476/j.pibb.2015.0078

* This work was supported by a grant from The National Natural Science Foundation of China(91430111, 61473232, 61401370, 61170134).
**Corresponding author.
Tel: 86-29-88431308, E-mail: zhangsw@nwpu.edu.cn
Received: March 23, 2015 Accepted: July 1, 2015